eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

UC | Universidad de Cantabria

Universidad de Oviedo

Máster Oficial en Economía: Instrumentos del Análisis Económico

Curso académico: 2023/2024

Trabajo de Fin de Máster

Inference in High-Dimensional two-way Panel Data Models

(Inferencia en datos de panel de alta dimensión con efectos fijos individuales y temporales)

Autora: Lindes Domínguez Díaz

Tutores: Juan Manuel Rodríguez Poo
Alexandra Pilar Soberón Vélez

Septiembre – 2024

DECLARACIÓN RESPONSABLE

Lindes Domínguez Díaz es la única responsable del contenido del Trabajo Fin de Master que se presenta. La Universidad de Cantabria, así como los profesores directores del mismo, no son responsables del contenido último de este Trabajo.

En tal sentido, la autora se hace responsable:
1. De la AUTORÍA Y ORIGINALIDAD del trabajo que se presenta.
2. De que los DATOS y PUBLICACIONES en los que se basa la información contenida en el trabajo, o que han tenido una influencia relevante en el mismo, han sido citados en el texto y en la lista de referencias bibliográficas.

Lindes Domínguez Díaz declara que el Trabajo Fin de Master tiene una extensión menor de 10.000 palabras, excluidas tablas, gráficos y bibliografía.

Fdo. Lindes Domínguez Díaz

Firmado por DOMINGUEZ DIAZ
LINDES - ***9845** el día
03/09/2024 con un certificado
emitido por AC FNMT Usuarios

Contents

Abstract

The aim of this master thesis is to obtain a $\sqrt{NT}$ - consistent and asymptotically normal estimation for a triangular simultaneous two-way high-dimensional panel data model. Estimating such models is challenging due to endogeneity from two sources: the correlation between variables in a two-way panel and the dependence between covariates and the error term. This thesis proposes a two-stage estimator where individual and time effects are first removed, followed by an instrumental variable estimation. Given high-dimensional data sets where the number of covariates exceeds the sample size, traditional methods fail. Instead, the Lasso method and its variants, Cluster-Lasso and Post-Lasso, are used for estimation, providing consistency and asymptotic normality under specific conditions.

Resumen

El objetivo de este trabajo de fin de máster es obtener una estimación $\sqrt{NT}$ - consistente y asintóticamente normal para un modelo de datos de panel triangular simultáneo de alta dimensión con efectos fijos individuales y temporales. La estimación de este tipo de modelos supone un reto debido a la endogeneidad de dos fuentes: la correlación entre variables en un panel con dos efectos fijos y la dependencia entre covariables y el término de error. Este trabajo propone un estimador en dos etapas en el que primero se eliminan los efectos individuales y temporales, seguido de una estimación de variables instrumentales. Ante conjuntos de datos de alta dimensión en los que el número de covariables supera el tamaño de la muestra, los métodos tradicionales fallan. En su lugar, el método Lasso y sus variantes, Cluster-Lasso y Post-Lasso, se utilizan para la estimación, proporcionando consistencia y normalidad asintótica bajo condiciones específicas.

Notation

Before start, let us define a number of mathematical concepts to facilitate the understanding of the following work. We have based ourselves on [10] and [4].

- $\#M$; given a set M, it means the cardinal of that set, that is, the number of elements that the set has.

- $\|\pi\|_0$; the l_0 norm returns the number of non-zero elements of the vector π, i.e., $\|\pi\|_0 = \#\{i : \pi_i \neq 0\}$.

- $\|\pi\|_1$; the l_1 norm is defined as $\|\pi\|_1 = \sum_{j=1}^{p} |\pi_j|$.

- $\|\pi\|_2$; the l_2 norm is defined as $\|\pi\|_2 = \sqrt{\sum_{j=1}^{p} \pi_j^2}$.

- $o(1),\ N \to \infty$; a function $v(N)$ that depends on N is $o(1)$, $N \to \infty$, if $\lim\limits_{N\to\infty} v(N) = 0$.

- $o_P(1)$; a sequence of random variables $\{X_N\}$ is said to be $o_P(1)$ if $X_N \xrightarrow{\text{p}} 0$.

- $O(1)$, $N \to \infty$; a function $v(N)$ that depends on N is $O(1)$, $N \to \infty$, if $|v(N)|$ remains bounded as $N \to \infty$.

- $O_P(1)$; a sequence of random variables $\{X_N\}$, with respective distribution functions $\{F_N\}$, is said to be bounded in probability ($O_P(1)$) if for every $\epsilon > 0$ there exist M_ϵ and N_ϵ such that $F_N(M_\epsilon) - F_N(M_\epsilon) > 1 - \epsilon \quad \text{all} \quad N > N_\epsilon$.

- $p \vee NT$; it means that it is taking the higher of the following values p and NT, i.e. $p \vee NT = \max(p, NT)$.

Chapter 1

Introduction

The aim of this master thesis dissertation is to obtain a $\sqrt{NT}$ - consistent and asymptotically normal estimation of a triangular simultaneous two-way high- dimensional panel data model. Traditionally, in a standard fully parameter setting, it has been challenging to estimate these type of models due to the double source of endogeneity they present. On one side, two-way panel data model exhibit the correlation of two random variables, one fixed effect that accounts for unknown individual effects that are usually correlated with the covariates and a time effect. On the other side, the second source of endogeneity comes from the dependence between the covariates of the structural equation and the idiosyncratic error term

The solution to these problems has been a two-stage estimator procedure where, in the first stage, a transformation that removes both individual and time effects is performed. Then, in a second stage, an instrumental variable estimator is derived over the transformed model. Under fairly general conditions the IV estimation is consistent and asymptotically normal. To obtain consistency, one of the conditions assumed in the reduced form equation is that the number of explanatory variables (instruments) is much smaller than the sample size and furthermore the matrix of explanatory variables is full row rank. This assumption is classic in linear regression models but nowadays, with the availability of big data, it might become rather unrealistic.

Increasingly, in Econometrics and Statistics we find high-dimensional data sets where the number of covariates is larger than the sample size. This situation raises problems of sparseness and violates a crucial assumption for the existence of different estimators in the linear regression model (full row rank condition). Indeed, in our econometric setting, the availability of a high-dimensional data set in the reduced form equation rends the IV estimation procedure asymptotically biased and therefore alternative estimation procedures are needed.

In the field of high-dimensional data there are numerous methods employed to achieve sparse models. Sparsity describes the presence of few non-zero values in a data set. Sparse models are easier to interpret as fewer variables make it easier to understand how the predictions are made. In our case we will rely on the Lasso method and two of its variants known as Cluster-Lasso and Post-Lasso, which are one of the key ideas of the article [2].

Under the presence of high-dimensional data sets in the reduced form equation, in this master thesis dissertation, we propose a Lasso and Cluster-Lasso type estimation to compute the reduced form equation and then a two-stage least squares estimation of the structural parameters. Under some conditions we show that the 2SLS of the structural parameters are $\sqrt{NT}$ - consistent and asymptotically normal.

This results follows the same lines as in [2] where a triangular simultaneous fixed effects panel data model with high-dimensional data was estimated. Unfortunately they did not include the time dimensional component in both the reduced form and the structural model.

In Section 1 we present the econometric model, in Section 2 we introduce the estimation procedure and we show the asymptotic properties, in Section 3 we conduct a Monte Carlo simulation study. Finally we conclude.

1.1 The econometric model

Let us consider the following econometric model:

$$y_{it} = \beta d_{it} + \alpha_i + \gamma_t + \epsilon_{it}, \tag{1.1}$$

$$\begin{aligned} d_{it} &= h(w_{it}) + f_i + l_t + u_{it}, \\ i &= 1, ..., N; \quad t = 1, ..., T, \end{aligned} \tag{1.2}$$

with $h(w_{it}) = z_{it}'\pi + r(w_{it})$ where $z_{it} = z(w_{it})$ a $1 \times p$ vector, can be any unknown transformation of the instrument w_{it} and $r(w_{it})$ is the remainder term. The term π is a $p \times 1$ vector, so the d_{it} term has dimension 1×1. In addition, we should mention that $p >> N$, i.e. the number of instruments is larger than the number of individuals observed. y_{it} is the outcome variable of interest, d_{it} represents an endogenous variable, $(\alpha_i,\ f_i)$ are the individual fixed effects and $(\gamma_t,\ l_t)$ are time fixed effects which represent the effects of the omitted variables that are specific to both individual units and time periods. ϵ_{it}, u_{it} are the idiosyncratic error terms.

Since d_{it} is defined as endogenous, $E[\epsilon_{it} d_{it} | \alpha_i, \gamma_t] \neq 0$ for $i = 1, ..., N$ and $t = 1, ..., T$. Endogeneity arises for a variety of reasons, three common sources of endogeneity stand out in the literature: omitted variables, a measurement error or simultaneity (see [14]).

In our model the source of endogeneity is simultaneity. Simultaneity arises when at least one of the explanatory variables d_{it} is determined simultaneously along with y_{it}. If, for example, d_k is determined partly as a function of y, d_k and ϵ_{it} are usually correlated. For example if y_{it} is the number of patients admitted to hospital for a specific disease and d_{it} is the number of doctors in charge of curing this disease, the number of doctors depends on the number of patients (see [14]).

Under the econometric problem (1.1) and (1.2) and having $E[\epsilon_{it} d_{it} | \alpha_i, \gamma_t] \neq 0$, clearly the OLS estimator of β is inconsistent. In addition, it should be mentioned that considering only the equation (1.1) and $E[\epsilon_{it} d_{it} | \alpha_i, \gamma_t] = 0$, the OLS estimation will be inconsistent due to the fixed effects (see [14]). The fixed effects problem can be solved with a transformation that will be explained in the next chapter.

The problem of correlation between variables and the error term can be solved using the method of instrumental variables (IV) (see [3]), more specifically, as mentioned in [2], we will use the two-stage least squares (2SLS) estimator. IVs are variables uncorrelated with the error term (disturbance term) but correlated with the endogenous explanatory variables, although they do not represent explanatory variables in the original regression model. The instrumental variables w_{it} appear in the second equation of our model which is often known as reduced form equation for the endogenous explanatory variable d_{it}.

It is also known as the linear projection of d_{it} on w_{it}.

In recent years, several researchers have studied the case in which the number of available variables for each observation is much larger than the number of observations which is known as high-dimensional data panels. The problem with these models is that we have too much information, i.e. too many variables, many of which are not relevant. This causes problems during estimation, making it necessary to use a method for selecting relevant variables.

In our model, the high-dimension is found in the second equation as the number of instruments is much larger than the sample size. Since, as previously mentioned, $p >> N$. As mentioned in [9], if in the case of high-dimensional an ordinary least squares estimation is performed, we find an overfitting of data which translates as a large variance and a very low bias implying that the estimate is not valid. The problem of overfitting is commonly known in the field of machine learning and statistics, which occurs when a model gives accurate predictions for training data but not for new data. This can occur for many reasons and one of them is that if the training data contains a lot of noise or irrelevant features, the model may try to fit the noise along with the underlying patterns. In our case we have a model with too many instrumental variables to estimate and many of them are not relevant. When we find ourselves in these situations, we resort to the so-called sparse or regularization methods, which use a penalty to reduce the number of explanatory variables in the model thus providing a solution to the bias-variance tradeoff. Ideally, a model should be chosen that accurately captures the regularities of the training data, but also generalises well to unseen data. Unfortunately, it is often impossible to do both at the same time. With this regularization methods we achieve a balance between variance and bias.

Another reason we cannot perform OLS when dealing with high-dimensional models is that the matrix containing the observations of the explanatory variables does not have full rank, i.e. since $p >> n$, the number of columns exceeds the number of rows, meaning there cannot be p linearly independent columns. Consequently, the rank of the matrix is at most N, which is less than p, making the matrix not full rank. A matrix that is not full rank is singular, meaning it has a determinant of zero and cannot be inverted. Since inverting that matrix is a crucial step in the OLS estimation, this singularity makes it impossible to compute the OLS solution using the standard formula.

Chapter 2

Estimation strategy and asymptotic properties

In this chapter, we present the estimation procedure that provides consistent estimation of the parameters of interest in high-dimensional panel data models under endogeneity. We will also explain some regularization methods such as Lasso, Cluster-Lasso and Post-Lasso. Finally we will present the regularity and asymptotic properties and some theorems related to them.

2.1 Transformation

When we have a model that includes only individual fixed effects, the most common transformation used to eliminate these fixed effects is the within transformation. In our case, in addition to an individual fixed effect, we also have a time fixed effect, as defined in (1.1). Therefore, our aim is to eliminate both fixed effects, for which we have used the following transformation:

$$\ddot{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} - \frac{1}{N}\sum_{i=1}^{N} y_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it} \tag{2.1}$$

Analogously, we define $\ddot{d}_{it}$, $\ddot{\epsilon}_{it}$ $\ddot{z}_{it}$, $\ddot{r}(w_{it})$ and $\ddot{u}_{it}$. After applying the above transformation our model is as follows

$$\begin{aligned} \ddot{y}_{it} &= \ddot{d}_{it}\beta + \ddot{\epsilon}_{it}, \\ \ddot{d}_{it} &= \ddot{h}(w_{it}) + \ddot{u}_{it} = \ddot{z}_{it}'\pi + \ddot{r}(w_{it}) + \ddot{u}_{it}. \end{aligned} \tag{2.2}$$

As the reader can see, applying this transformation we have obtained a model without fixed effects. If $E[\ddot{\epsilon}_{it}\ddot{d}_{it}] = 0$ then the OLS estimator of β is consistent and asymptotically normal. Unfortunately, under endogeneity of $\ddot{d}_{it}$, $E[\ddot{\epsilon}_{it}\ddot{d}_{it}] \neq 0$, the OLS estimator of β presents an asymptotic bias. In the following we present the estimation technique that enables us to obtain consistent estimates of β in the presence of endogenous covariates d_{it}. Previously, we will explain some of the regularization methods that are part of the estimation process.

2.2 Regularization Methods

As mentioned above, in many occasions, we encounter panel data sets in which the number of explanatory variables per observation is very large, i.e. $p >> N$. In these cases, regularization methods are recommended to reduce the number of explanatory variables in the model as we see in [9]. These methods have the advantage that they simultaneously perform the selection of variables and the estimation of the coefficients of the selected variables.

Regularization methods are a class of techniques widely used in statistics and machine learning to address high-dimensional problems, where the number of explanatory variables or features p is large compared to the number of observations N. These methods are essential when dealing with models where the features may outnumber the observations, which can lead to problems such as overfitting, multicollinearity, and instability of the estimated coefficients.

The key idea of the regularization methods is to incorporate a penalty term to the cost function that penalises models with many large parameters. By minimising this new function, some coefficients become zero, simplifying the model and thus achieving a sparse model. The choice of the penalty parameter determines the intensity of this penalty term and is crucial to obtain a good model. It is therefore necessary to use appropriate methods to estimate it from the data.

In our case we find a high-dimensional problem in the reduced form (1.2) since the length of the vector z_{it} is p , the size of our sample is N and $p >> N$. Therefore, in order to estimate it consistently we have to apply regularization methods.

2.2.1 Ridge Regularization

In 1970 Robert Hoerl and Kennard introduced Ridge regression which is one of the best known regularization methods (see [7]). With Ridge regularization we estimate the coefficients $\pi_0, \pi_1, ..., \pi_p$ that minimise the following expression:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\ddot{d}_{it} - \ddot{z}'_{it}\pi)^2 + \frac{\lambda}{NT}\sum_{j=1}^{p}\pi_j^2 \tag{2.3}$$

where $\lambda \geq 0$ is a parameter and the second term, $\frac{\lambda}{NT}\sum_{j=1}^{p}\pi_j^2$, called as shrinkage penalty, makes small when $\pi_0, \pi_1, ..., \pi_p$ are close to zero.

This method helped a lot in improving the simulation of high-dimensional models and has served as an inspiration for many other regularization methods. However, this regularization model uses the l_2 norm which does not achieve a sparse model. This is because using the l_2 norm we cannot get any coefficient to be exactly zero. As λ increases, the coefficients tend to zero, but they never become zero and this creates a problem for interpretation because all variables both relevant and irrelevant, are included in the final model. This is the main disadvantage of this method of regularization because even if the irrelevant variables have a very small coefficient, they will always have a non-zero value, making it difficult to interpret the results.

When we perform least squares estimation, we seeks coefficients that minimize only the first term of the expression above $\frac{1}{NT}\sum_{i=1}^{N}(\ddot{d}_{it} - \ddot{z}'_{it}\pi)^2$. But with Ridge reguralization, we introduce the penalty term and with that we can control the variance. As λ increases, the variance decreases but, at the same time the bias increases. When the penalty term has no effect, $\lambda = 0$, Ridge regression will produce least squares estimation and in this case the variance is high and there is no bias.

In this regularization, the way in which the penalty term is defined causes all variables to be included in the model. As λ increases, the coefficients towards zero,but they never become zero and this creates a problem for interpretation because all variables, relevant and irrelevant, are included in the final model.

2.2.2 Lasso Regularization

In 1996 Robert Tibshirani introduced an improved alternative to the Ridge regression, (see [11]), the Lasso (Least Absolute Shrinkage and Selection Operator). The Lasso coefficient estimate π_L is the solution to the penalised minimization problem defined as follows:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\ddot{d}_{it} - \ddot{z}'_{it}\pi)^2 + \frac{\lambda}{NT}\sum_{j=1}^{p}|\pi_j|. \tag{2.4}$$

It uses the l_1 norm, which allows some estimated coefficients in the Lasso regularization to be exactly zero when the parameter λ is sufficiently large, i.e. to obtain a sparse model at the same time as the parameter estimation. This is one of the advantages of the Lasso because in the Ridge method all variables are included in the model, but in the Lasso a variable selection is performed. As a result of that, models produced by Lasso regularization are easier to interpret than those generated from Ridge regression.

2.2.3 Cluster-Lasso Regularization

Another estimation method for obtaining sparse models is the Cluster-Lasso, an extension of the Lasso technique based on the Group-Lasso introduced by Yuan and Lin in 2006 (see [15]). While traditional Lasso selects variables individually, these techniques group related variables together, which facilitates the interpretation of the results and reduces the dimension of the model.

In many practical problems, such as genetics, variables (such as genes) are organised into groups based on their function or location. Selecting whole groups rather than individual variables improves the interpretability of the model and the robustness of the estimates. In addition, variables within a group are often highly correlated. Traditional Lasso can have difficulty selecting relevant variables in the presence of multicollinearity, as it tends to select only one variable from a group of correlated variables. Cluster-Lasso addresses this problem by allowing the selection of complete groups of variables, which improves the stability of the estimates.

The main difference between Group-Lasso and Cluster-Lasso lies in how the groups of variables are defined. In Group-Lasso, the groups are previously established by the researcher, whereas in Cluster-Lasso, the groups are discovered automatically during the model fitting process. This is particularly useful when no prior information on the structure of the data is available. Group-Lasso is simpler and more computationally efficient, but less flexible. On the other hand, Cluster-Lasso is more flexible and can discover hidden structures in the data, but is computationally more expensive.

As defined in [2], the Cluster-Lasso coefficient estimate $\hat{\pi}_{CL}$ minimizes the following optimization problem:

$$\frac{1}{NT}\sum_{i=1}^{N}(\ddot{d}_{it} - \ddot{z}_{it}'\pi)^2 + \frac{\lambda}{NT}\sum_{j=1}^{p}\hat{\phi}_j|\pi_j|. \tag{2.5}$$

In this case in order to solve this optimization problem we need to assign a value to both the penalty level, λ, and covariates penalty loadings $\{\hat{\phi}_j\}_{j=1}^{p}$.

According to [2], it is important that the regularization event is verified for the Cluster-

Lasso estimates to be correct. The condition is as follows

$$\frac{\lambda\hat{\phi}_j}{NT} \geq 2c\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{z}_{itj}\ddot{u}_{it}\right| \quad \text{for each} \quad 1 \leq j \leq p, \tag{2.6}$$

where $c > 1$ is a constant slack parameter. It should be noted that $\dfrac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{z}_{itj}\ddot{u}_{it}$ is a natural measure for the noise in estimating π_j.

Therefore, the regularization event corresponds to the choice of a penalty parameter that is large enough to dominate the noise in the model coefficient estimates. We can see that the regularization event causes all coefficients whose amplitudes are not large enough relative to the sampling noise to be set exactly to zero in the Lasso solution. This property makes Lasso-based methods attractive for prediction and variable selection in order to achieve sparse models, where many model parameters can be assumed to be zero, and it is desirable to exclude from the model all variables that cannot be reliably determined to have a strong distribution.

It is necessary that condition (2.6) is fulfilled with high probability. To achieve the condition, according to the authors in [2] the intuition for suitable choices can be seen by considering the following equality for ϕ_j, where

$$\phi_j^2 = \frac{1}{NT}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\ddot{z}_{itj}\ddot{u}_{it}\right)^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\ddot{z}_{itj}\ddot{z}_{it'j}\ddot{u}_{itj}\ddot{u}_{it'j},$$

captures the sampling variability in learning about coefficient π_j. It can be observed that the values of ϕ_j depend on the unobservable error term, $\ddot{u}_{it}$. Therefore, in order to estimate ϕ_j, we must first estimate $\hat{\ddot{u}}_{it}$ and then calculate $\hat{\phi}_j$. To obtain $\hat{\ddot{u}}_{it}$ we use use an algorithm designed by [2] that we will explain later in the simulations chapter. Thus, after applying the algorithm to calculate $\hat{\ddot{u}}_{it}$, our optimal $\hat{\phi}_j$ is as follows

$$\hat{\phi}_j^2 = \frac{1}{NT}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\ddot{z}_{itj}\hat{\ddot{u}}_{it}\right)^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\ddot{z}_{itj}\ddot{z}_{it'j}\hat{\ddot{u}}_{it}\hat{\ddot{u}}_{it'}.$$

It is important that the penalty loading verifies the following

$$\begin{aligned}&\ell\phi_j \leq \hat{\phi}_j \leq r\phi_j, \text{with probability close to } 1,\\ &\text{for some}\, r \leq C < \infty \,\text{and}\, \ell \to 1, \text{uniformly for}\, j = 1, ..., p.\end{aligned} \tag{2.7}$$

According to [2], under the above condition and setting :

$$\lambda = 2c\sqrt{NT}\Phi^{-1}(1 - \zeta/2p),$$

where $c > 1$ is a constant, $\zeta = o(1)$ and Φ^{-1} is the inverse cumulative normal distribution function. The regularization event is verified with probability close to one.

2.2.4 Post-Lasso/Post-Cluster-Lasso Regularization

The Post-Lasso method is a statistical technique that is used as a kind of 'second step' after applying the Lasso. Post-Lasso takes the model obtained with Lasso and refines it to obtain better statistical properties. The Post-Cluster-Lasso method works in exactly the same way as Post-Lasso but using the model obtained with the Cluster-Lasso method. For this reason and to simplify the notation, we will use Post-Lasso to refer to both methods.

This estimator is simply ordinary least squares applied to the data after removing the regressors that were not selected by Cluster-Lasso, which we define as $\hat{P}_\pi = \{j : \hat{\pi}_j \neq 0\}$. The Post-Lasso estimator $\hat{\pi}_{PL}$ minimizes

$$\frac{1}{NT}\sum_{i=1}^{N}(\ddot{d}_{it} - \ddot{z}'_{it}\pi)^2, \tag{2.8}$$

only on the set $\hat{P}_\pi$.

2.2.5 Data driven selection of the λ parameter

In all of this regularization methods is critical to select a good value of the parameter λ. The parameter λ in these regularisation methods mentioned above acts as a regulator that allows us to control the balance between bias and variance. The Bias-Variance tradeoff is a fundamental concept in statistics and machine learning that describes the relationship between two sources of error that affect the performance of a predictive model: bias and variance.

Bias refers to the systematic difference between the model's predictions and the actual values it is trying to predict. High bias implies that the model is making simplifying assumptions about the data, which can lead to underfitting. Underfitting occurs when the model is too simple to capture the underlying patterns in the data.

Variance refers to the sensitivity of the model to fluctuations in the training data. A model with high variance fits the training data very closely, capturing both real patterns and noise. This can lead to overfitting, where the model performs extremely well on the training set, but fails to generalise to new data.

Bias-variance tradeoff describes the tradeoff needed between these two types of errors to minimise the total prediction error of a model. In general models with high bias tend to be simpler, but may not capture the complexity of the data (underfitting). Models with high variance tend to be more complex, capturing both real patterns and noise in the training data (overfitting).

The goal is to find a middle ground where the model is complex enough to capture the true patterns in the data, but not so complex as to be influenced by noise. This balance results in a model that generalises well to new data.

As we increase λ, the penalty term becomes stronger and more coefficients go to zero. This introduces bias in the model, as we are forcing some coefficients to be zero, even if they have a real effect on the dependent variable. by reducing the complexity of the model (fewer variables), we reduce the variance. A model with fewer variables is less likely to over-fit the training data and will therefore generalise better to new data.

If λ is very small, the penalty term is weak and the model will closely resemble a standard linear regression model. This can lead to overfitting and high variance. Conversely, if λ is very large, the penalty term is strong and many coefficients will shrink to zero. This can lead to an underfitted model and high bias.

It is important to choose a parameter that provides a balance between small variance and controlled bias. Although a small variance is important, a very large bias can produce inaccurate estimation. Therefore, the confidence in the results decreases, thus affecting the interpretation of these estimations.

The authors in [2] use for the Lasso and Cluster-Lasso methods the previously defined λ. However, there are many other methods to obtain the most suitable value. The following are some of the most well-known methods.

Cross-validation (CV)

One of the most commonly used methods for the estimation of the λ term is the k-fold cross-validation, which consists on dividing the data set into k subsets or 'folds' of the same size. The idea is to train the model k times, using $k-1$ folds for training and the remaining fold for validation, each of the folds used exactly once as the testing set. The mean square error, MSE, is then calculated for each of the folds (see [9]).

In the specific case of the selection of the λ parameter, the steps to be followed to obtain the best value with CV are the following:

1. Select the number of folds k.

2. Divide the data into training and test sets.

3. Define a grid of values for λ.

4. For each λ calculate the validation MSE within each fold.

5. For each λ calculate the overall cross-validation MSE.

6. Selecting the λ that minimises the cross-validation MSE.

Based on [6], we denote by $M = \{(\ddot{d}_{it}, \ddot{z}_{it}), \quad i = 1, .., N \quad \text{and} \quad t = 1, ..., T\}$ for the set of all the data, and the training and test set by $M - M^v$ and M^v respectively, for $v = 1, ..., p$. Each M^v is made up of a fixed proportion of randomly selected elements of M. The value of λ obtained by k-folds CV shall be the λ that minimises the following expression:

$$CV(\lambda) = \frac{1}{k} \sum_{v=1}^{k} \frac{1}{\# M^v} \sum_{(\ddot{d}_{it}, \ddot{z}_{it}) \in M^v}^{N} (\ddot{d}_{it} - \ddot{z}_{it}' \hat{\pi}_v(\lambda))^2, \tag{2.9}$$

By evaluating the performance of the model on different subsets of the data, we can find the value of λ that best fits our data.

AIC and BIC

According to [12], although it has been found that the CV method is a better way to choose the λ parameter, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are methods which can be used to obtain the value of the λ term, but their use is restricted to estimates that are linear in their parameters.

The optimum λ for the AIC method is the one that minimises the following expression

$$AIC = -2 \cdot log(\hat{\mathcal{L}}) + 2 \cdot p, \tag{2.10}$$

where $\hat{\mathcal{L}}$ is the maximum value of the likelihood function for the model and p is the number of regressors (in our case the length of the vector z_{it}).

In the case of the BIC method, the optimal λ is the one that minimises the following expression

$$BIC = -2 \cdot log(\hat{\mathcal{L}}) + p \cdot log(NT). \tag{2.11}$$

The expression is very similar to the previous one but in this case the logarithm of the number of observations is used.

2.3 Two-stage least squares estimator

In this section we will explain the procedure we have used to estimate the full model formed by the two equations (1.1) and (1.2). According to [14], the two-stage least squares (2SLS) estimator is the most efficient IV estimator. The first thing to do is to apply the transformation explained in the previous section to remove the fixed effects from our model. Once the fixed effects have been eliminated, we can start applying the 2SLS to our transformed model (2.2).

In the first stage, we estimate the reduced form model, as mentioned above, we have a high-dimensional problem since $p >> N$, where p is the length of the vector z_{it} and N the size of our sample. Therefore, in order to estimate consistently it we have to apply regularization methods. For this purpose we apply the Lasso or Cluster-Lasso method to estimate $\hat{\pi}$ and simultaneously obtain a sparse model. In this way we manage to reduce the size of the vector of instruments z_{it}. Once one of these two methods has been applied, we are able to define the subset $\hat{P}_\pi = \{j : \hat{\pi}_j \neq 0\}$ and then apply the Post-Lasso to the reduced form on that subset.

The second stage of the 2SLS estimator consists of estimating the first equation of the model and obtaining the value of $\hat{\beta}$. The second stage of the 2SLS estimator is essentially an ordinary least squares (OLS) regression, but using the predicted values of the first stage as the explanatory variable. After performing OLS, the estimation of β is

$$\hat{\beta} = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{d}_{it} \hat{\ddot{H}}_{it} \right)^{-1} \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\ddot{H}}_{it} \ddot{y}_{it} \right), \tag{2.12}$$

where $\hat{\ddot{H}}_{it} = \ddot{z}'_{it} \hat{\pi}_{PL}$.

This estimator is also known as the IV estimator (see [14]). In the following, we summarise what has been explained in this section in an algorithm that will be used in subsequent simulations.

Algorithm to calculate 2SLS estimator

We will now describe the steps necessary to obtain the estimator presented above.

1. Apply the transformation (2.1) to the two equations (1.1) and (1.2).

2. Use the Lasso or Cluster-Lasso technique to estimate the transformed reduced form equation.

3. Apply the Post-Lasso method to the transformed reduced form equation only with the variables selected in step 2 (i.e. the variables that the Lasso or Cluster-Lasso

considers non-zero).

4. Calculate $\ddot{d}_{it}$ with coefficients estimated with Post-Lasso, i.e. $\hat{\ddot{H}}_{it} = \ddot{z}_{it}' \hat{\pi}_{PL}$
5. Apply OLS to the first equation transformed using $\hat{\ddot{H}}_{it}$ instead of $\ddot{d}_{it}$ to obtain $\hat{\beta}$. Note that this is equivalent to estimate β in (1.1) through an IV estimator using $\hat{\ddot{H}}_{it}$ as instruments.

2.4 Regularity and asymptotic properties

In this section, we present the conditions under which the Cluster-Lasso and Post-Lasso methods are valid and allows us to make inference. Therefore, we analyse the main asymptotic features of the proposed estimator. In order to present these conditions, we consider the following additive fixed effects model, which for simplicity of notation is identical to the reduced form previously presented.

$$d_{it} = h(w_{it}) + f_i + l_t + u_{it} \quad i = 1, ..., N \quad t = 1, ..., T,$$

where $E[u_{it}|w_{i1}, ...w_{iT}, f_i, l_1, ..., l_T] = 0$. The time invariant individual specific heterogeneity is represented by f_i. The term l_t represents the time fixed effects which are identical for all individuals in the sample. Note that both fixed effects are correlated with w_{it} and we assume that the sequence $\{d_{it}, w_{it}\}_{t=1}^{T}$ is independent and identically distributed (i.i.d.) across i but does not impose any restrictions on the dependence within individuals. Furthermore, in this case we consider

$$h(w_{it}) = z_{it}'\pi + r(w_{it}).$$

Before explaining the conditions, we define the index of information, which was used in [2] allowing for within-individual dependence. It is defined as

$$\iota_T := T \min_{1 \le j \le p} \frac{E\left[\dfrac{1}{T}\sum_{t=1}^{T} \ddot{z}_{itj}^2 \ddot{u}_{itj}^2\right]}{E\left[\dfrac{1}{T}(\sum_{t=1}^{T} \ddot{z}_{itj} \ddot{u}_{itj})^2\right]} \tag{2.13}$$

We have two extreme cases of this index

- $\iota_T = 1$; in this case we have no information and corresponds to perfect dependence within the cluster i.
- $\iota_T = T$; we have maximal information that corresponds to perfect independence within the cluster i.

Apart from these two cases, there are many more in between. Now, having defined this index, let us enumerate the conditions under which the estimator performs well and returns sparse estimations with good predictive properties and convergence rate.

2.4.1 Regularity conditions and results

1. **Condition ASM** (Approximately Sparse Model)
   A function $h(w_{it})$ is said to be well-approximate by a linear combination of transformations $z_{it} = Z_{NT}(w_{it})$, where Z_{NT} is a measurable map and z_{it} is a vector of length p, if for each i and t

   $$h(w_{it}) = z_{it}'\pi + r(w_{it}),$$

   where the remainder term $r(w_{it})$ and the coefficient β satisfy

   $$\|\pi\|_0 \leq s = o(N\iota_T) \quad \text{and} \quad \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} r(w_{it})^2\right]^{1/2} \leq B_s = O_P(\sqrt{s/N\iota_T}).$$

   This condition indicates that the number of non-zero coefficients of the vector β, i.e. the number of predictor variables selected by the model is less than or equal to s. The condition $o(N\iota_T)$ means that the number of non-zero coefficients is of order less than $N\iota_T$. This means that the number of non-zero coefficients in the model grows at a rate smaller than N, as the sample size increases.

   The remainder inequality relates the l_2 norm to B_s and establishes that the l_2 norm of the error term is upper bounded by B_s, and in turn, B_s is upper bounded by $\sqrt{s/N\iota_T}$ with high probability as the sample size increases. Therefore, we can see how the l_2 norm of the approximation error is bounded by a value that depends both on the number of non-zero predictors, the sample size and also on the index of information.

   In summary, these restrictions suggest that as the sample size and the index of information increase, the number of significant predictors does not grow excessively fast and also the mean square error decreases as the sample size increases. This helps to maintain a balance between the number of predictors in the model and the amount of information available in the data, which can be crucial to avoid overfitting or unreliable models.

   The following condition allows us to control the behaviour of the Gram matrix, which is a $p \times p$ matrix of the covariances between the variables.

   $$\ddot{M} = \{M_{jk}\}_{j,k=1}^{p}, \qquad M_{jk} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{z}_{itj}\ddot{z}_{itk}.$$

   In standard regression analysis, when the sample size is larger than the number of variables, the Gram matrix must have full rank. In high-dimensional panel models, usually the Gram matrix will be singular, because we have more variables than observations, and for the matrix to be of full rank the sample size should be the same as the number of variables, i.e. $N = p$. However, for Lasso and Cluster-Lasso to work properly, it only requires good behaviour of certain moduli of continuity of $\ddot{M}$.

Before setting out the condition we define the minimal and maximal i-sparse eigenvalues of this matrix as follows

$$\varphi_{min}(i)(\ddot{M}) := \min_{\delta \in \triangle(i)} \delta' \ddot{M} \delta \quad \text{and} \quad \varphi_{max}(i)(\ddot{M}) := \max_{\delta \in \triangle(i)} \delta' \ddot{M} \delta,$$

where

$$\triangle(i) = \left\{ \delta \in \mathbb{R}^p : \|\delta\|_0 \leq i, \quad \|\delta\|_2 = 1 \right\},$$

is the i-sparse subset of a unit sphere.

2. **Condition SE** (Sparse Eigenvalues)
For any $C > 0$, there exist $k', k'' \in \mathbb{R}$, where $0 < k' < k'' < \infty$, which may depend on C but do not depend on N, such that with probability approaching one, as $N \longrightarrow \infty$, $k' \leq \varphi_{min}(Cs)(\ddot{M}) \leq \varphi_{max}(Cs)(\ddot{M}) \leq k''$

The above condition refers to the fact that only certain $Cs \times Cs$ sub-matrices smaller than the original matrix, the Gram matrix, are required to be well-behaved for the estimator to work properly. The size of this submatrices depends on a constant C and s, the number of non-zero predictors.

3. **Condition R** (Regularity Conditions) Suppose that for the data $\{d_{it}, w_{it}\}$ the following conditions are satisfied with z_{it} defined as in the first condition, that is $z_{it} = Z_{NT}(w_{it})$,with probability close to 1:

   - $\left(\frac{1}{T} \sum_{t=1}^{T} E[\ddot{z}_{itj}^2 \ddot{u}_{itj}^2]\right) + \left(\frac{1}{T} \sum_{t=1}^{T} E[\ddot{z}_{itj}^2 \ddot{u}_{itj}^2]\right)^{-1} = O(1)$, as $T \longrightarrow \infty$

     The previous equality implies that the relationship between the explanatory variables and the error terms should neither increase nor decrease excessively as the number of time periods increases.

   - $1 \leq \max_{1 \leq j \leq p} \phi_j / \min_{1 \leq j \leq p} \phi_j = O(1)$.

     This condition ensures that the relation between the maximum of the penalties and the minimum is bounded between 1 and a constant.

- $1 \leq \max_{1\leq j\leq p} \varpi_j \sqrt{E\phi_j^2} = O(1)$ where $\varpi_j = \left(E\left[\left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \ddot{z}_{itj}\ddot{u}_{itj} \right|^3 \right] \right)^{1/3}.$
- $\log^3(p) = o(NT)$ and $\mathrm{slog}(p \vee NT) = o(N\iota_T)$.

  The first part of this affirmation states that the cube of the logarithm of the number of predictors, p, grows at a slower rate than the total sample size multiplied by the number of time periods. This indicates that the growth of p is relatively slow compared to the sample size and the number of time periods.
  The second part implies that the product of s and the logarithm of the maximum between p and NT grows at a slower rate than $N\iota_T$.
- $\max_{1\leq j\leq p} \left| \phi_j - \sqrt{E\phi_j^2} \right| / \sqrt{E\phi_j^2} = o(1).$

All of these conditions are the same that have been used by the authors in [2], where only a within transformation has been used. Remember that our transformation is the following

$$\ddot{d}_{it} = d_{it} - \frac{1}{T}\sum_{t=1}^{T} d_{it} - \frac{1}{N}\sum_{i=1}^{N} d_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} d_{it}.$$

Assuming $E[d_{it}] < \infty$, $E[z_{it}] < \infty$, $E[u_{it}] = 0$ and that they are all i.i.d random variables across i and that are stationary along t. For fixed T, applying the law of large numbers (see [13]), $\frac{1}{N}\sum_{i=1}^{N} d_{it} \xrightarrow{\text{a.s.}} E[d_{it}]$.

Also by the law of large numbers and by the stationary, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} d_{it} \xrightarrow{\text{a.s.}} \frac{1}{T}\sum_{t=1}^{T} E[d_{it}] = E[d_{it}]$.

Therefore, asymptotically our transformation is

$$\ddot{d}_{it} = d_{it} - \frac{1}{T}\sum_{t=1}^{T} d_{it} + o_P(1),$$

that is commonly known as within transformation. The same applies to z_{it} and u_{it} as explained for the variable d_{it}. Consequently, the conditions ASM, SE and R are valid for our transformation and attain favourable performance bounds. With all the above conditions, we can establish the asymptotic Cluster-Lasso performance bounds that are collected in the following theorem.

Theorem 1 *(Selection properties of Cluster-Lasso and Post-Lasso models) Consider a sequence of probability laws $\{P_{n,T}\}$ for which $\{(d_{it}, w_{it}, z_{it})\}_{t=1}^{T} \sim P_{N,T}$ i.i.d across i for which $N \to \infty$, T fixed. Assume that Conditions ASM, SE and R hold for probability* *measure $P = P_{N,T}$ induced by $P_{N,T}$. Consider a Lasso or Cluster-Lasso estimator defined in previous sections with the penalties generated by the algorithm. The subset $\hat{P}$ satisfies with probability close to 1, $\hat{s} = |\hat{P}| \leq Ks$ for some $K > 0$ which does not depend on the size of the sample, N. Then the Lasso or Cluster-Lasso estimator and the Post-Lasso estimator verify the following conditions*

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\ddot{z}_{it}'\hat{\pi} - \ddot{z}_{it}'\pi)^2 = O_P\left(\frac{s\log(p \vee NT)}{N\iota_T}\right),$$

$$\|\hat{\pi} - \pi\|_2 = O_P\left(\sqrt{\frac{s\log(p \vee NT)}{N\iota_T}}\right),$$

$$\|\hat{\pi} - \pi\|_1 = O_P\left(\sqrt{\frac{s^2\log(p \vee NT)}{N\iota_T}}\right).$$

Note that Theorem 1 only ensures asymptotic bounds for the Cluster-Lasso type estimators obtained for the reduced form in equation (1.2).

2.4.2 Asymptotic conditions and results

In this section we will provide conditions to ensure that $\hat{\beta}$ is a $\sqrt{NT}$ - consistent estimator and moreover, we will also give a consistent estimator for the variance-covariance matrix. Let

$$\begin{aligned} y_{it} &= \beta d_{it} + \alpha_i + \gamma_t + \epsilon_{it}, \\ d_{it} &= h(w_{it}) + f_i + l_t + u_{it}, \\ i &= 1, ..., N; \quad t = 1, ..., T. \end{aligned}$$

It is important to mention that the above conditions (regularity conditions) must also be fulfilled in order to achieve this. To simplify the notation we first define the following quantities:

For any arbitrary random variables, $A = \{A_{it}\}_{i\leq N, t\leq T}$

$$\phi^2(A) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_{it}\right)^2,$$

$$\varpi(A) = E\left[\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} A_{it}^{3}\right|\right]^{1/3},$$

$$\iota_T(A) = T\frac{E\left[\frac{1}{T}\sum_{t=1}^{T} A_{it}^2\right]}{E[\phi^2(A)]},$$

and with that we can define

$$\phi_j^2 = \phi^2(\{\ddot{z}_{itj}\ddot{u}_{it}\}), \quad \varpi_j = \varpi(\{\ddot{z}_{itj}\ddot{u}_{it}\}), \quad \iota_T = \min_{1\leq j\leq p} \iota_T(\{\ddot{z}_{itj}\ddot{u}_{it}\})$$

$$\phi_H^2 = \phi^2(\{\ddot{H}_{it}\ddot{\epsilon}_{it}\}), \quad \varpi_H = \varpi(\{\ddot{H}_{it}\ddot{\epsilon}_{it}\}), \quad \iota_T^H = \iota_T(\{\ddot{H}_{it}\ddot{\epsilon}_{it}\})$$

$$\phi_{zjd}^2 = \phi^2(\{\ddot{z}_{itj}\ddot{d}_{it}\}), \quad \varpi_{zjd} = \varpi(\{\ddot{z}_{itj}\ddot{d}_{it}\}), \quad \iota_T^{zjd} = \iota_T(\{\ddot{z}_{itj}\ddot{d}_{it}\})$$

$$\phi_{zj\epsilon}^2 = \phi^2(\{\ddot{z}_{itj}\ddot{\epsilon}_{it}\}), \quad \varpi_{zj\epsilon} = \varpi(\{\ddot{z}_{itj}\ddot{\epsilon}_{it}\}), \quad \iota_T^{zj\epsilon} = \iota_T(\{\ddot{z}_{itj}\ddot{\epsilon}_{it}\})$$

1. **Condition SMIV**

   - $\frac{1}{T}\sum_{t=1}^{T} E\left[\ddot{H}_{it}^2\right]$, $\frac{1}{T}\sum_{t=1}^{T} E\left[\ddot{\epsilon}_{it}^2\ddot{H}_{it}^2\right]$, $E\left[\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{d}_{it}^2\right)^2\right]$ are uniformly bounded in N and T above and far from zero.

   - Higher order moments are bounded, i.e., $E\left[\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{\epsilon}_{it}^2\right)^q\right] = O(1)$ for some $q > 4$.

   - $\dfrac{\varpi_D}{\sqrt{E\phi_H^2}} = O(1)$, $\max\limits_{1\leq j\leq p}\dfrac{\varpi_{zj\epsilon}}{\sqrt{E\phi_{zj\epsilon}^2}} = O(1)$.

   - $\max\limits_{j}\dfrac{\iota_T^{zj\epsilon}}{T}\phi_{z_j\epsilon}^2 = O_P(1)$, $\dfrac{\phi_{dD}^2}{T} = O_P(1)$, $\max\limits_{j}\dfrac{\phi_{zjd}^2}{T} = O_P(1)$.

   - $\dfrac{s^2\log^2(p\vee NT)}{N\iota T}\max\left\{1, \max_{1\leq j\leq p}\dfrac{\iota_T^H}{\iota_T^{zj\epsilon}}\right\} = o(1)$.

   - $\dfrac{\iota_T^H}{\iota_T}N^{2/q}\dfrac{s\log(p\vee NT)}{N} = o(1)$.

   These conditions guarantee that the parameter β would be strongly identified if $\ddot{H}_{it}$ could be observed. It also implies that the use of a small number of variables in z_{it} is sufficient to identify β accurately,

Before enunciating the theorem, we define an estimator of the asymptotic variance of $\hat{\beta}$, which will be necessary to perform inference for the parameter β after appropriately rescaling, as

$$\hat{V} = Q^{-1} \left[\frac{1}{NT} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{t'=1}^{T} \ddot{d}_{it}\ddot{d}_{it'}\hat{\ddot{\epsilon}}_{it}\hat{\ddot{\epsilon}}_{it'} \right] Q^{-1},$$

where $\hat{Q} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{d}_{it}\hat{\ddot{H}}_{it}$.

Theorem 2 *(Estimation and Inference in IV Models) Let be a sequence of probability laws $\{P_{N,T}\}$ for which $\{(y_{it}, d_{it}, z_{it})\}_{t=1}^{T} \sim P_{N,T}$ i.i.d across i for which $N \to \infty$, T fixed and for which the instrumental variable model holds. Assume that the regularity and asymptotic conditions are satisfied, then the estimator of β verify*

$$\sqrt{N\iota_T^D}\, V^{-1/2}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, 1),$$

and

$$V - \frac{\iota_T^H}{T}\hat{V} \xrightarrow{P} 0.$$

where V is defined as

$$V = \frac{\iota_T^H}{T} Q^{-1} \Omega Q^{-1},$$

and

$$Q = \frac{1}{T} \sum_{t=1}^{T} E\left[\ddot{H}_{it}^2\right], \qquad \Omega = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{t'=1}^{T} E\left[\ddot{H}_{it}\ddot{H}_{it'}\ddot{\epsilon}_{it}\ddot{\epsilon}_{it'}\right].$$

This theorem verifies that the estimator of β constructed with instruments selected by Lasso or Cluster-Lasso in a linear IV model with fixed effects, is both consistent and asymptotically normal. Furthermore,the theorem states that we can use $\hat{V}$ to perform valid inference for β after instrument selection. This inference remains valid uniformly across a broad range of data-generating processes, including scenarios where perfect instrument selection is unattainable.

Chapter 3

Simulation results

This section analyses the small sample properties of the estimator proposed in the previous section. To this end, based on [2], we have performed Monte Carlo simulations with both methods, Lasso and Cluster-Lasso. For both methods we have used the same data set in order to see the performance of both and compare the results obtained.

We generate data from the following model, which has already been defined above:

$$\begin{aligned} y_{it} &= \beta d_{it} + \alpha_i + \gamma_t + \epsilon_{it} \\ d_{it} &= z_{it}'\pi + f_i + l_t + u_{it}. \end{aligned}$$

We define the error terms as

$$\begin{aligned} \epsilon_{it} &= \rho_\epsilon \epsilon_{it-1} + \nu_{1,it}, \\ u_{it} &= \rho_u u_{it-1} + \nu_{2,it}, \end{aligned}$$

where

$$\begin{pmatrix} \nu_{1,it} \\ \nu_{2,it} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_\nu \\ \rho_\nu & 1 \end{pmatrix} \right) \text{ i.i.d.}$$

We generate time fixed effects for $t = 1, ..., T$ and for $i = 1, ..., N$ we define the individual fixed effects. In addition we set $\alpha_i = f_i$ and $\gamma_t = l_t$.

$$\gamma_t \sim \mathcal{N}(0, 1), \qquad \alpha_i \sim \mathcal{N}\left(0, \frac{4}{T}\right).$$

Our instruments are defined as follows

$$z_{itj} = \alpha_i + \rho_z z_{i(t-1)j} + \phi_{itj}, \quad t > 1, \tag{3.1}$$

where ϕ_{itj} is a $\mathcal{N}(0,1)$. Finally, the structure of the coefficients on the instruments, π. The coefficient vector π is defined as

$$\pi_j = (-1)^{j-1}\frac{1}{\sqrt{s}}\mathbf{1}_{\{j\leq s\}} + \frac{1}{j^2}\mathbf{1}_{\{j>s\}}, \quad s = \left\lfloor \frac{1}{2}N^{1/3} \right\rfloor,$$

for $1 \leq j \leq p$ and $\left\lfloor \frac{1}{2}N^{1/3} \right\rfloor$ is the integer part of $\frac{1}{2}N^{1/3}$. We consider two numbers of instruments, $p_1 = N \times (T-2)$ and $p_2 = N \times (T+2)$. We have carried out 500 simulations for different sample sizes, $N = 25$, 50 and 100 all with $T = 10$.

When implementing the Lasso and Post-Lasso methods it is common to use the cross validation technique to obtain the optimal λ value for the data set. In the case of the λ based on [1] and [2], taking c=1.1 and $\zeta = 0.1/\log(p \vee NT)$, the value obtained is too large, which results in no variable being selected in the first step of the 2SLS, preventing the process from continuing. This could be because our transformation is not the same as the one used by the authors in [1] and [2]. Although we have shown that they behave asymptotically the same, for the sample sizes used, where our sample is finite, their λ parameter does not provide an optimal result.

To show the difference between the λ parameter defined by the authors and the λ parameter obtained with the cross-validation technique we have performed an iteration and the results are as follows:
Cross-Validation: 0.121
Optimum λ according to [1] and [2]: 139.535

Seeing the difference, for our simulations we have used the λ provided by the cross-validation technique. In addition, we have performed simulations for two other values of λ which are 0.01 and 0.3, one value larger and one value smaller than the one obtained with CV to see how the regularization methods behave with changes in the λ parameter.

In this case, it is not necessary to split the data into training and test sets when the goal is to estimate a specific parameter, such as β, and calculate its bias and RMSE with respect to a known value, because using all available data allows a more accurate estimation of the parameter and a more direct assessment of the bias. Partitioning is fundamental in machine learning to assess the predictive performance of the model on unseen data and to prevent overfitting. As in our case we want to observe how well our model estimates, we do not need to split the sample. The table below shows the results of our simulations with the Lasso method, in the table we show the bias and the RMSE obtained for the different sample sizes in order to evaluate the robustness of our model.

Table 3.1: Simulations Results (Lasso)

		$p_1 = N \times (T-2)$		$p_2 = N \times (T+2)$	
		Bias	**RMSE**	**Bias**	**RMSE**
$\lambda = 0.01$	$N = 25$	0.377	0.393	0.383	0.396
	$N = 50$	0.369	0.376	0.374	0.381
	$N = 100$	0.370	0.374	0.379	0.382
$\lambda = CV$	$N = 25$	0.370	0.387	0.368	0.386
	$N = 50$	0.362	0.371	0.373	0.381
	$N = 100$	0.365	0.370	0.370	0.375
$\lambda = 0.3$	$N = 25$	0.372	0.394	0.379	0.396
	$N = 50$	0.364	0.375	0.369	0.379
	$N = 100$	0.363	0.369	0.374	0.381

As we can see in the table, the Lasso method works optimally for the λ calculated with CV, but not for the fixed λ of 0.01 and 0.3. Lasso with cross-validation allows for more flexible adaptation to the data by automatically selecting the optimal λ value for each data set. As the sample size increases, the prediction error estimate becomes more accurate, which in turn leads to a more reliable selection of λ. This results in a decrease in the RMSE, as the model better fits the underlying relationships in the data. In contrast, a fixed value of λ imposes a rigid constraint on the model, regardless of the variability in the data. Therefore, as the sample size increases, the model with fixed λ may not improve significantly or even worsen its performance, especially if the value of λ is not optimal for the new sample size as we see in our case.

As mentioned in Chapter 1, the Cluster-Lasso has the advantage of adding $\{\hat{\phi}_j\}_{j=1}^{p}$ to the traditional Lasso. To calculate the values of these parameters we have based ourselves on the algorithm proposed by the authors in [2] but applied to our model.

Algorithm to calculate $\hat{\phi}_j$

Define for $j = 1, ...p$ an initial $\hat{\phi}_j$ value

$$\text{Initial:} \quad \hat{\phi}_j = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\ddot{z}_{itj}\ddot{z}_{it'j}\ddot{d}_{it}\ddot{d}_{it'}}, \tag{3.2}$$

and a refined one

$$\text{Refined:} \quad \hat{\phi}_j = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\ddot{z}_{itj}\ddot{z}_{it'j}\hat{\ddot{u}}_{it}\hat{\ddot{u}}_{it'}}. \tag{3.3}$$

In both cases, the λ value is the previously defined. We denote by $K \geq 1$ the number of iterations.

Steps of the algorithm:

1. With the values of $\ddot{z}_{itj}$ and $\ddot{d}_{it}$ applied in (3.2) we obtain ϕ_{j0}.
2. Using $\hat{\phi}_{j0}$ in the optimization problem (2.5) or (2.8) and solving this problem we obtain $\hat{\pi}_j$. Then, we calculate $\hat{\ddot{u}}_{it} = \ddot{d}_{it} - \ddot{z}'_{itj}\hat{\pi}_j$ for $i = 1, ..., N$ and $t = 1, ..., T$.
3. Update the penalty loadings using the $\hat{\ddot{u}}_{it}$ calculated in the step 2 applied in (3.3).
4. Use the new $\hat{\phi}_j$ in one of the optimization problem and calculate $\hat{\pi}_j$.
5. Repeat the step 4 $K - 1$ times.

In practise this algorithm is performed for K number of iterations. The more iterations the more accurate our estimation of $\ddot{u}_{it}$ will be. But it is important to take into account the computational cost and find a balance between calculated cost and accuracy. We have iterated the algorithm 12 times and after obtaining the refined $\{\hat{\phi}_j\}_{j=1}^{p}$ we have applied the Cluster-Lasso model to our data.

As with the Lasso method, we performed the simulations for three different values of the λ parameter. We have used the λ provided by the cross-validation technique and two fixed values that are the same as in the Lasso method, 0.01 and 0.3. The following table shows the results obtained in the simulations.

Table 3.2: Simulations Results (Cluster-Lasso)

		$p_1 = N \times (T-2)$		$p_2 = N \times (T+2)$	
		Bias	**RMSE**	**Bias**	**RMSE**
$\lambda = 0.01$	$N = 25$	0.376	0.393	0.383	0.397
	$N = 50$	0.369	0.376	0.374	0.381
	$N = 100$	0.370	0.374	0.379	0.382
$\lambda = CV$	$N = 25$	0.369	0.387	0.368	0.385
	$N = 50$	0.363	0.371	0.373	0.381
	$N = 100$	0.365	0.370	0.371	0.375
$\lambda = 0.3$	$N = 25$	0.371	0.392	0.378	0.396
	$N = 50$	0.362	0.372	0.369	0.379
	$N = 100$	0.361	0.368	0.376	0.382

In the table we can see that for the λ value obtained with the cross-validation method it works well because the RMSE decreases as we increase the sample size. For the two cases where we have set the λ value to 0.01 and 0.3, the method stops working optimally and the RMSE is not decreasing. It is important to mention that the values obtained for the Cluster-Lasso method are for most cases identical to the results obtained with the Lasso method.

3.1 Conclusions

The results in the two tables show that the two methods work very similarly. Therefore, for the econometric model studied in this paper, it would be indifferent to use either the Lasso or the Cluster-Lasso method. The simulations have shown that in our case it was not appropriate to give a fixed value for the λ term because we do not get results that improve as the sample size increases.

From the results obtained with the fixed values of λ, it is worth noting that there is practically no difference in the results between those obtained with the value 0.01 and the value 0.3, which indicates that the increase in λ does not give better results and we only obtain good results by looking for a λ with CV or with some method that obtains this parameter based on the data.

However, it should also be noted that the computational cost of simulations with a fixed λ and with a λ obtained with CV is very different. Implementing an algorithm to obtain the λ value has a very high computational cost, but to obtain adequate results it is necessary. Therefore, we can conclude that despite the computational cost, the Lasso and Cluster-Lasso together with a λ obtained with CV works adequately.

Based on the simulations performed, and for the case where the λ parameter is obtained with CV, we observe that both the bias and the RMSE of the estimators decrease as we increase the sample size. It is true that there is a time when the bias increases, but compared to the smaller sample size it does decrease. Since the RMSE decreases as the sample size increases, the variance of the estimator is decreasing. This is consistent with the theory that an estimator is asymptotically normal, since the variance decreases as the sample increases, bringing the distribution of the estimator closer to a normal one. In absolute terms (looking at the first size $N = 25$ and the last $N = 100$) we see that the bias decreases, we can conclude that our estimator is consistent. Although we have not directly calculated the consistency of the variance-covariance matrix estimator, given that the estimators of the structural parameters are consistent and the variance-covariance matrix is calculated based on these estimators, it is reasonable to expect that it is also consistent.

Bibliography

[1] Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369–2429.

[2] Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. Journal of Business and Economic Statistics, 34(4), 590–605.

[3] Bowden, R. J. and Turkington, D. A. (1990). Instrumental variables. Cambridge University Press.

[4] Boyd, S., and Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press.

[5] Giraud, C. (2021). Introduction to high-dimensional statistics (2nd ed.). Chapman and Hall/CRC.

[6] González Vidal, A.(2015). Selección de variables: Una revisión de métodos existentes. [Master Thesis Dissertation, Universidade de Coruña]

[7] Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

[8] Hsiao, C. (2014). Analysis of panel data (3rd ed.). Econometric Society Monographs.

[9] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.

[10] Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. Hoboken, New Jersey: Wiley.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

[12] Van Le, C. (2020). How to Choose Tuning Parameters in Lasso and Ridge Regression? Asian Journal of Economics and Banking, 4(1), 61–76.

[13] White, H. (1984). Asymptotic theory for econometricians. Academic Press.

[14] Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data (2nd ed.). MIT Press.

[15] Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 49-67.

Appendix

R code to implement 2SLS estimator

Below is the R studio code that we have implemented in order to perform the 2SLS estimator using the Lasso method.

The first stage of the two-stage least squares:

```
#Lasso
lasso.mod <- glmnet(z_transformed,d_transformed,alpha=1,lambda
   =0.3)
coef_lasso_df <- as.data.frame(as.matrix(coef_lasso))
#Non-zero coefficients selected with Lasso
non_zero_coef <- coef_lasso_df[coef_lasso_df != 0, , drop=FALSE]
non_zero_variables <- rownames(non_zero_coef)[rownames(non_zero_
   coef) != "(Intercept)"]
z_transformed_ filtered <- z_transformed[, non_zero_variables,
   drop=FALSE]
#Post-Lasso
post_lasso<-lm(d_transformed~z_transformed_ filtered)
# Calculate d with coefficients estimated with Post-Lasso
d_estimated <- coef(post_lasso)[1] + as.matrix(z_transformed_
   filtered) %*% coef(post_lasso)[-1]
```

The second stage of the two-stage least squares is the following:

```
# Ordinary least squares estimation
ols <- lm(y_transformed~d_estimated)
# Beta estimated
beta_estimated <- coef(ols)["d_estimated"]
```

Following is the R code to implement the 2SLS using the Cluster-Lasso method.

The first stage of the two-stage least squares:

```
#Cluster-Lasso
cluster.lasso.mod <- glmnet(z_transformed,d_transformed,penalty.
   factor = phi_refined,alpha=1,lambda=0.3)
coef_cluster_lasso<- coef(cluster.lasso.mod)
coef_cluster_lasso_df <- as.data.frame(as.matrix(coef_cluster_
   lasso))
#Non-zero coefficients selected with Cluster-Lasso
non_zero_coef <- coef_cluster_lasso_df[coef_cluster_lasso_df !=
   0, , drop=FALSE]
non_zero_variables <- rownames(non_zero_coef)[rownames(non_zero_
   coef) != "(Intercept)"]
z_transformed_filtered <- z_transformed[, non_zero_variables,
   drop=FALSE]
#Post-Lasso
post_lasso<-lm(d_transformed~z_transformed_filtered)
# Calculate d with coefficients estimated with Post-Lasso
d_estimated <- coef(post_lasso)[1] + as.matrix(z_transformed_
   filtered) %*% coef(post_lasso)[-1]
```

The second stage of the two-stage least squares is the following:

```
# Ordinary least squares estimation
ols <- lm(y_transformed~d_estimated)
# Beta estimated
beta_estimated <- coef(ols)["d_estimated"]
```

Following is the R code to implement the 2SLS using the CV method.

The first stage of the two-stage least squares:

```
# CV
cv<-cv.glmnet(z_transformed,d_transformed,alpha=1)
bestlam<- cv$lambda.min
#Cluster-Lasso
cluster.lasso.mod <- glmnet(z_transformed,d_transformed,penalty.
   factor = phi_refined,alpha=1,lambda=bestlam)
coef_cluster_lasso<- coef(cluster.lasso.mod)
coef_cluster_lasso_df <- as.data.frame(as.matrix(coef_cluster_
   lasso))
#Non-zero coefficients selected with Cluster-Lasso
non_zero_coef <- coef_cluster_lasso_df[coef_cluster_lasso_df !=
   0, , drop=FALSE]
non_zero_variables <- rownames(non_zero_coef)[rownames(non_zero_
   coef) != "(Intercept)"]
z_transformed_filtered <- z_transformed[, non_zero_variables,
   drop=FALSE]
#Post-Lasso
post_lasso<-lm(d_transformed~z_transformed_filtered)
# Calculate d with coefficients estimated with Post-Lasso
d_estimated <- coef(post_lasso)[1] + as.matrix(z_transformed_
   filtered) %*% coef(post_lasso)[-1]
```

The second stage of the two-stage least squares is the following:

```
# Ordinary least squares estimation
ols <- lm(y_transformed~d_estimated)
# Beta estimated
beta_estimated <- coef(ols)["d_estimated"]
```